TINTRÍ

Building Tintri VMstore

TECHNICAL WHITE PAPER

By Tintri Architects Mark Gritter & Ed Lee

何故Tintri VMstoreを開発したのか?

バーチャリゼーションはデータセンターをVM形態へ変貌させたことで大きな変化を遂げました。バーチャルインフラストラクチャー上のアプリケーションは、この進化をもって初めて論理的なオブジェクトとして捉えられるようになりました。物理的なマシンでは大変難しいことですが、バーチャルアプリケーションは複製、再設定、再構築、分析、管理が容易に行えます。仮想化はサーバ統合のメリットだけでなく、データセンター管理における構築、運用の負担を軽減します。

残念ながら、バーチャルマシン(VMs)のストレージは急速にボトルネックとなってしまいました。そして、この問題を回避できなければVM構造への完全な脱却は成立しません。VMは汎用的なハードウェアリソースの集合体上で稼働させることができ、そのCPUやメモリの使用状況を簡単にモニタリングしたり、変更したりすることができます。ところが、従来ストレージにおいては問題が発生します。仮想化技術が登場する前に開発された従来のストレージは、個別の設定が必要だからです。このミスマッチがバーチャルアプリケーションを物理的な領域へと引き戻してしまうのです。

Tintriはこれらの制限を乗り越えたソリューションを提案します。VM専用に開発され、VMストレージの問題に注力したTintri VMstoreはその他の仮想インフラと同等レベルの管理が可能になります。

図1:Tintriは複雑な設計項目を必要としません

Tintriはフラッシュテクノロジー、ファイルシステムのアーキテクチャ、ユーザーインターフェースの利点を取り込み、バーチャルアプリケーションをシンプルかつ効果的なストレージにします。

LUNs≠VMs

バーチャリゼーションとの互換性のため、シェアドストレージの（SAN（Fibre ChannelやiSCSI）もNAS（NFS）も）導入は急速に進みました。しかし、従来のネットワークストレージ製品の仮想化には幾つかの障害がありました。VMsにとって意味を成さないLUN、ボリューム、tierなどを管理するためです。従来ストレージはVMs程詳細なモニタリングをしたり、スナップショットを取ったり、ポリシーを設定したり、データを複製したりすることができません。

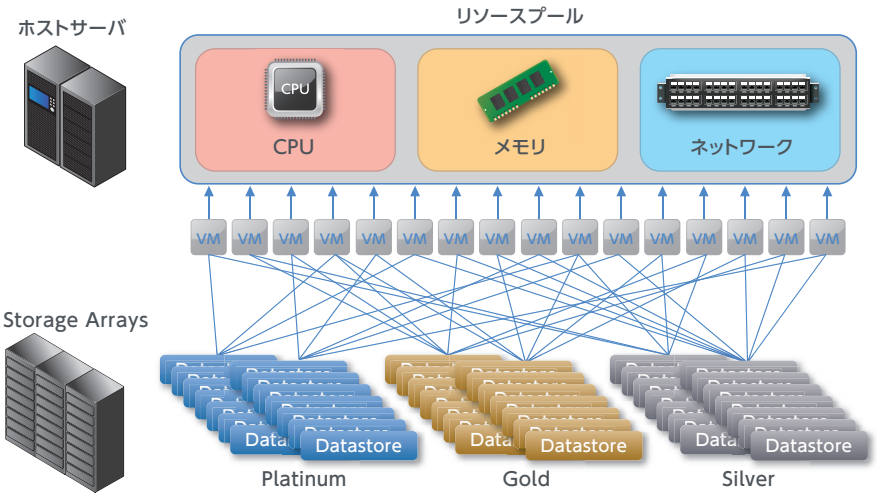

図 2：複雑な現行ストレージ

このミスマッチにより、コストと複雑さが増します。新規のVMインスタンスは特定のLUNもしくはボリュームに紐付けられなければなりません。I/Oの要件とVMの挙動が正しく理解されない場合には、ここからトライアンドエラーの作業を開始しなければなりません。ストレージの管理者とVMの管理者はお互いに、全アプリケーションのためのスペース（領域）が確保されているかを確認し、適切なI/Oパフォーマンスと、予想される容量を調整しなければなりません。
通常、複数のVMsはマッピングの複雑さと容量のオーバーヘッドを削減するためにLUNと同等のボリュームを使用します。しかし、これによってI/Oパフォーマンスの問題が更に複雑化します。ストレージを中心に据えたパフォーマンスデータを基準とするなら、管理者はどのVMsに影響が出るのか、どのVMsが負荷を生じさせているのかを予測しなければなりません。

ストレージマネジメントのオーバーヘッドの削減を目的とした自動ティアリングですら間違った構造で稼働しているのです。VM毎、もしくはバーチャルディスクレベル毎の挙動をレポートできなければ、先進的なストレージ技術は不透明な技術のままです。ハイパーバイザーからもたらされるシンプルなVMモデルの代わりに、従来のストレージは途方もない数の選択肢やインターフェースで応答してきます。

こういった状況下においては、VMsの設定、管理、調整は複雑でコストが高く、仮想化の導入を制限してしまいます。言い換えるならば、従来のネットワークストレージに多数のアプリケーションを仮想化しても、費用対効果が見込めないということです。

VMを意識したストレージ

Tintriは、VM専用に開発されたストレージにおいて、一般的なストレージより仮想環境の方が効果的であるというシンプルな思想の基に設立されました。Tintri VMstoreはバーチャリゼーションとストレージ両分野のエキスパートによって 1 から開発され、フラッシュ、マルチコアCPU、インライン重複除外・圧縮といった最新のテクノロジーを搭載しています。

Tintri VMstoreはLUNやボリューム単位ではなく、VMsやバーチャルディスク単位で管理されています。Tintriのファイルシステムは VM環境の要求を満たし、VMsに適した機能を提供するために 1 から開発されました。Tintriの革新的なフラッシュ技術の使用により、シンプルかつコスト効果を高く、条件の厳しいアプリケーションをバーチャル環境に移動させるために必要なパフォーマンスが提供可能になりました。

この技術革新により、ストレージの管理は個別に設定するコンポーネントから、総体的なVMsへと変化しました。現在のコンピューターインフラに多く見られるのは、パフォーマンス、管理、コストなどの問題です。Tintriはこれらの問題によって仮想化が妨げられるのを阻止します。Tintriがより良いVMsストレージシステムを開発しようとする狙いは、全く新しいタイプの製品を生み出し、VMsにおける現行ストレージの制限を解消するためです。

どうやって実現するのか?

管理インターフェース

TintriがVMsに焦点を当てていることは、管理インターフェースに最もよく現れています。管理インターフェースでは、LUN、ボリューム、ファイル単位でなく、VMs単位でデータが表示されます。インターフェースの全オブジェクトはVM管理者にとって理解しやすく、直観的に表示されます。そのGUIは、ストレージを直観的に管理できるほどのシンプルさを持ち、なおかつ多数のVMsストレージをきめ細やかに管理可能なほどの洗練さも持ち合わせています。

Tintri VMstore上の各VMのパフォーマンス統計値は個別に取得され、インターフェースに表示されます。特定のVMやバーチャルディスクのユーティリゼーションやパフォーマンスの変化も容易に把握することができます。従来からある計測項目(レーテンシー、使用容量、スループット)に加え、Tintriの管理インターフェースではVMs毎、かつ全システムに渡って主なパフォーマンスを使用しているVMがどれなのか(フラッシュ容量、CPU、ディスクのスループット)を提示することができます。VMsやバーチャルディスク毎に的確な情報を提示することで、パフォーマンスのトラブルシューティングやキャパシティプランニングを容易にします。

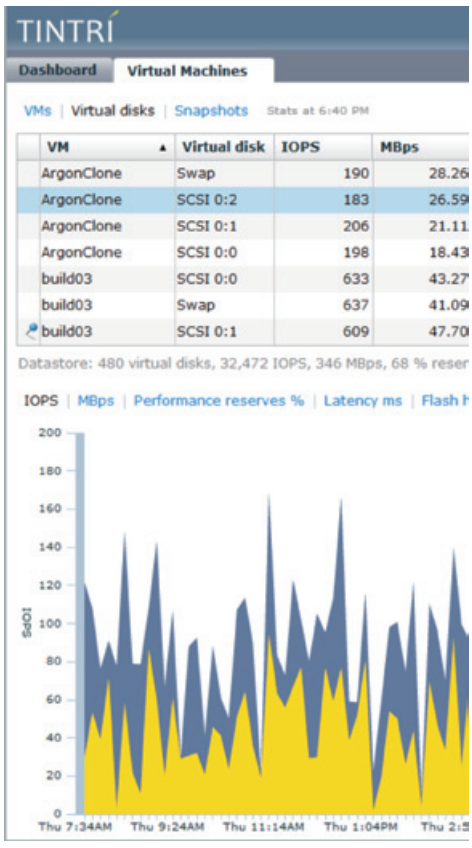

図 3：VM毎の各項目の詳細

コアファイルシステム

VMに特化した管理インターフェースの構築は、ただ見栄えの良いGUIを作るのとは遥かに異なる作業が必要です。裏で稼働しているストレージシステムは、VMレベルのパフォーマンス、キャパシティ・モニタリング、スナップショット、QoS管理、レプリケーションなどを理解し、サポートするように設計されています。

VMsに特化することで、Tintriは一般的なストレージシステムに必要とされるVMには不要なマッピングや複雑さを除外することができます。ストレージ層での意思決定が可能になり、一般的なストレージシステムより高度な自動化と最適化を実現することができます。シンプルな構造とインターフェースを持ち合わせた先進的なアーキテクチャが、更なる自動化と最適化に繋がります。

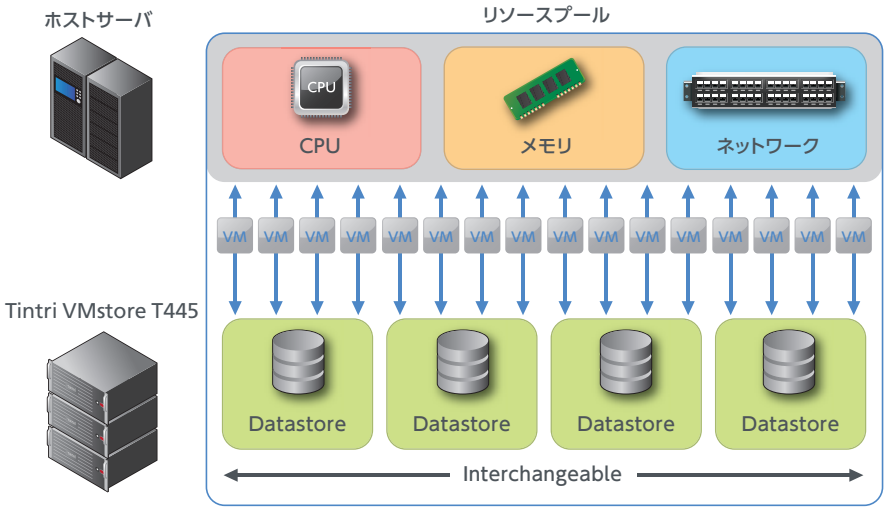

図4：よりシンプルな構造とインターフェース

例えば、異なるプロパティ（設定情報）を持った複数のユニットにストレージを細分化する必要がなくなります。それぞれのVMsストレージは、個別の利用目的に合わせて最適化されていきます。

TintriのVMに特化したコンセプトを利用すれば、更に条件の厳しい運用環境も最適化することができます。何十億というファイルを扱うようなシステムを開発する代わりに、Tintriは数百から数千のVMを問題なく扱えるシステムを開発し、VM毎にモニタリング、スナップショット、QoS、レプリケーションできるようにしました。

ソフトウェアスタックのどのシングルレベルでも、QoSを完全に実行することは困難です。何故なら、QoSが最も効果を発揮するのがハードウェア（例、VM毎のフラッシュ使用率）に近い部分なのか、クライアント（例、VM毎のIOPS）に近い部分なのか、システムの中間部分なのかがわからないからです。TintriのVMに特化したアプローチの場合には、適切なVM毎のQoSを実行することができます。VMwareのようなベンダーが提供するQoSは、ハイパーバイザーレベルで幾つか問題を発生させ、ストレージのサブシステムに利用するには困難で、非効率なQoSです。TintriのQoS機能は、このようなQoSと非常に上手く協業することができます。Tintriのファイルシステムは、ハイパーバイザーが各バーチャルディスクにどのようにアクセスしているかを把握し、その内容に沿ってリソースの保存場所を最適化することができます。

MLCフラッシュ

マルチレベルセル (MLC) フラッシュは、1 台のアプライアンスに何百、あるいは何千というVMを載せる際に発生する、多くのランダムI/O要求を処理するためにTintriが利用している重要な技術です。MLC SSDは高速なランダムI/Oパフォーマンスを提供できます。しかも、単に磁気ディスクの代わりとして利用するにはもったいない特異性を保持しています。しかし、そういった潜在的な利点を引き出すには、技術的に解決しなければならない課題があります。

最も明確な課題はフラッシュキャパシティのコストです。MLCフラッシュはSLCフラッシュの 1/2 から 1/4 程の価格ですが、SATAの 20 倍以上も高価な物です。フラッシュのキャパシティを一定に保つために、Tintriは高速なインライン重複除外・圧縮を、フラッシュとSATA間を自動的に移動するハイブリッドのファイルシステムと組み合わせています。既存のVMsを複製 (クローン) した多くのVMsを稼働させているバーチャル環境、もしくは同じOSやアプリケーションがインストールされたバーチャル環境において、インラインの重複除外・圧縮機能は大変効果的です。同様に、ハイブリッドのファイルシステムは、決まったサイズ以上のストレージを抱えたVMsの環境と同じくらい円滑にSATA上にしかないVMsを処理することができます。TintriのVMstoreにおいては、フラッシュはキャッシュでも、事前に設定された個別ストレージtierでもありません。フラッシュは、高いリードパフォーマンスが最大限に活用されるように設計されています。

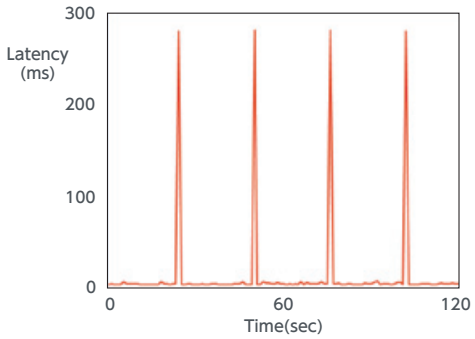

図 5：一般的なSSDのlatency spike

フラッシュにおけるその他の課題は、write amplificationとlatency spikeです。フラッシュコントローラーは、一般的なI/Oリクエストよりも大きなブロックサイズで消去を実行します。これを実現するために、ほとんどのフラッシュデバイスはストレージを管理するためのメタデータを保持しています。大体の場合、フラッシュデバイスはメタデータやガベージコレクトセルを保持するために余分なインターナルI/Oを生成します。この問題はwrite amplificationと呼ばれ、パフォーマンスを大幅に低下させる場合があります。メタデータアップデートのオーバーヘッドも同様にlatency spikeを頻繁に引き起こし (図 5)、I/Oリクエストの処理に通常の 100 倍もしくは 1000 倍以上の時間がかかってしまうことになります。Tintriは特許申請中の技術を利用してwrite amplificationとlatency spikeを防止します。

従来のフラッシュの技術では、MLCをベースにしたシステムは耐久性と信頼性においては脆弱である場合があります。特に、MLCセルは 5,000 ～ 10,000 回程度しか上書きができません。Tintriは複製、圧縮、高度なトランザクションやガベージコレクト技術を、フラッシュデバイスを継続的にモニタリングする技術と組み合わせ、MLCフラッシュの寿命を延ばします。加えて、フラッシュのRAID6 を採用し、新しいストレージデバイスから発生し得る影響を除外します。Tintri VMstoreはMLCフラッシュの強みを活かし、弱点を取り除くことでエンタープライズアプリケーションに適した、信頼性の高い、耐久性のあるストレージシステムを提供します。

まとめ

従来のネットワークストレージの複雑さ、パフォーマンス、コストの問題によって仮想化環境の導入が制限されています。バーチャリゼーションとストレージのエキスパートによって、VMs のためだけに 1 から開発された Tintri VMstore はこれらの問題を解決します。

Tintri が VM に特化したコンセプトを打ち出したことにより、構築が容易になり、コアファイルシステムが普及し、その結果を管理インターフェースで確認することができるようになりました。これにより、多くの利点が生じます。

Tintri VMstore により、企業インフラの仮想化を妨げるシステムの複雑さ、パフォーマンス、そしてコストの問題を解決することができます。

発売元

ノックス株式会社
www.nox.co.jp

本　　社 〒152-0023 東京都目黒区八雲2-23-13　Tel. 03-5731-5551　Fax. 03-5731-5552
西日本支社 〒533-0033 大阪市東淀川区東中島1-17-5 Tel. 06-4809-5544　Fax. 06-4809-5547

●本製品に関するお問い合わせ：営業本部
●メールでのお問い合わせ：tintri@nox.co.jp

お問い合わせ先

201107-A